Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.). We:

- propose a novel generative adversarial network that is able to solve occlusions in pedestrian images by hallucinating the missing parts while keeping both the appearance and the background coherent;
- devise a new way for synthetically generating occlusion pairs that result in more realistic images when compared to other methods previously employed;
- propose a method for conditioning the occluded body part restoration on pedestrian attributes and consequently improving the generation process;
- provide a large scale CG dataset for pedestrian attribute recognition in crowded areas;
- conduct an ablation study in order to clarify and highlight the solutions adopted in our work.

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Can Adversarial Networks Hallucinate Occluded People With a Plausible Aspect?

Federico Fulgeri[a], Matteo Fabbri[a,**], Stefano Alletto[a], Simone Calderara[a], Rita Cucchiara[a]

[a]*University of Modena and Reggio Emilia, Via P. Vivarelli 10, Modena, 41125, Italy*

ABSTRACT

When you see a person in a crowd, occluded by other persons, you miss visual information that can be used to recognize, re-identify or simply classify him or her. You can imagine its appearance given your experience, nothing more. Similarly AI solutions can try to hallucinate missing information with specific deep learning architectures, suitably trained with people with and without occlusions. The goal of this work is to generate a complete image of a person, given an occluded version in input, that should be a) without occlusion b) similar at pixel level to a completely visible people shape c) capable to conserve similar visual attributes (e.g. male/female) of the original one. For the purpose we propose a new approach by integrating the state-of-the-art of neural network architectures, namely U-nets and GANs, as well as discriminative attribute classification nets, with an architecture specifically designed to de-occlude people shapes. The network is trained to optimize a Loss function which could take into account the aforementioned objectives. As well we propose two datasets for testing our solution: the first one, occluded RAP, created automatically by occluding real shapes of the RAP dataset from Li et al. (2016) (which collects also attributes of the people aspect); the second is a large synthetic dataset AiC, generated in computer graphics with data extracted by the GTA video game, that contains 3D data of occluded objects by construction. Results are impressive and outperform any other previous proposal. This result could be an initial step to many further researches to recognize people and their behavior in an open crowded world.

1. Introduction

While recent efforts in people detection, recognition and tracking enabled a plethora of video-surveillance applications, e.g. people identification, pose estimation and action analysis, as in Ma et al. (2017); Riza Alp Guler (2018); Herath et al. (2017), occlusions are still an open problem. The occlusion issue is well known in the people detection and tracking literature and generally affects any intelligent video surveillance system, but it is debatable whether a real solution to the problem could exist effectively. In fact, whenever an occlusion occurs we observe a removal of information from the observed scene. The occluded portion of an object, indeed, becomes unknown and, in a Parmenidian sense, it does not exist until it can be observed. For this motivation most of the literature focused on counteracting the phenomenon conveying occlusion robustness to either detection, tracking or re-id systems as in Zhuo et al. (2018); Subramaniam et al. (2016); Pan and Hu (2007); Wang et al. (2018b); Coppi et al. (2016). In the matter of fact, recovering the image content from an occlusion is feasible only in the case where the target has been previously observed e.g. in a video stream. This is the approach followed also by many tracking solutions which memorize several detected appearance of the person, to discard occlusions as "less frequent accidents" w.r.t. the normal visible appearance. Nevertheless, leveraging on the generative capabilities of GANs in Goodfellow et al. (2014), we aim at demonstrating that it is indeed possible to hallucinate a plausible representation of the occluded content even when it has never been previously observed, i.e. in single images.

Following on our previous work on the topic (Fabbri et al. (2017)) in this paper we introduce a novel generative adversarial network that leverages the generative power of GANs for hallucinating the occluded portion of the image without any guidance of an attention mechanism that could provide instance level information about the occluding person. The proposed

**Corresponding author: Tel.: +39-340-577-8217
e-mail: matteo.fabbri@unimore.it (Matteo Fabbri)

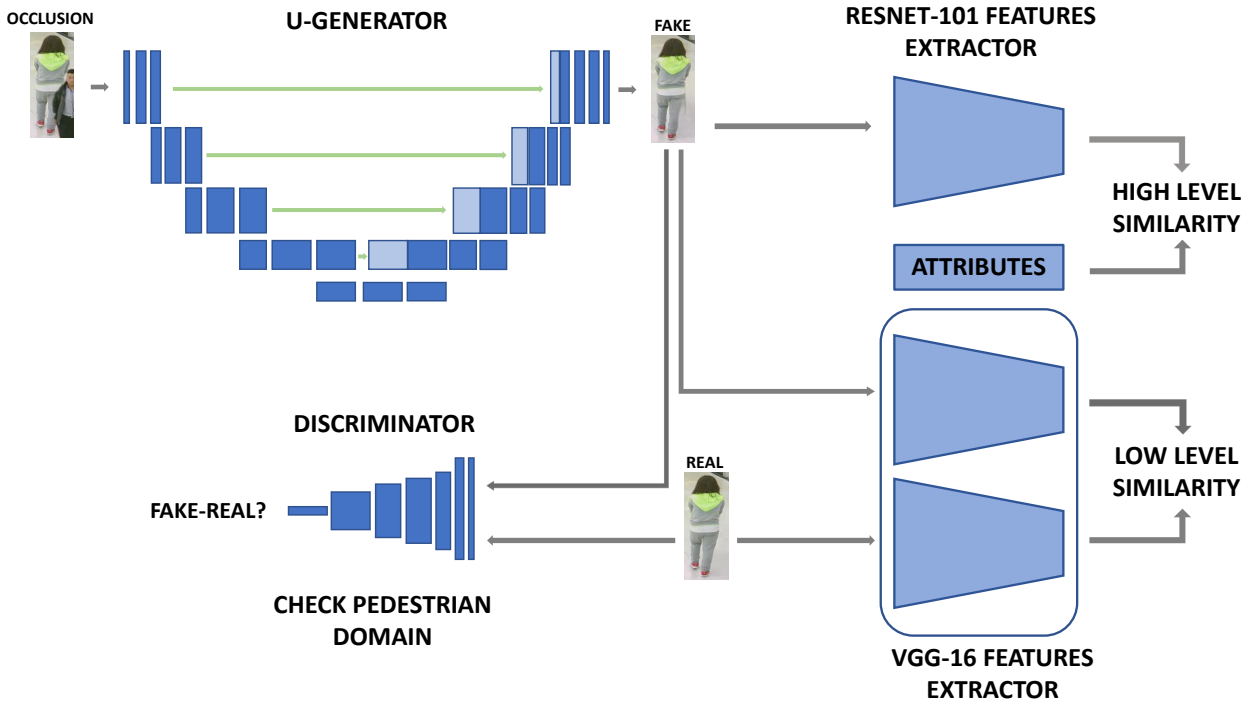

Fig. 1. A schematic representation of our method. On the left part is depicted the Adversarial Networks components, with the U-net generator above and the discriminator above. In the right part of the image is presented RESNET-101 used as high level features extractor (pedestrian attributes) and VGG-16 used as medium-low level features extractor.

solution aims at generating or reconstructing the image of a person which could be plausible in many senses: a) similar to images of real people, observed in the training dataset; b) acceptable at pixel level as a person shape; c) capable to preserve similar visual attributes of the ground truth de-occluded image. This is carried out by exploiting solutions for attribute calssifications (e.g. male/female, young/old, with/without trousers, etc.) and integrating them in a U-net like generative and adversarial architecture.

Another major problem that arises when dealing with occlusions, through learning-based solutions, is the lack of large-scale datasets providing realistic occluded and non-occluded pairs of images. Most of the proposed solution in literature, like Fabbri et al. (2017); Ouyang et al. (2016); op het Veld et al. (2015), paste together different people detections, or manually add random objects or textures to a non-occluded image. These processes ultimately fail to generate realistic data and are thus a liability when employed for training a neural network that aims at resolving the occlusion while keeping the rest of the image coherent (e.g. the background) and preserving the person's attributes. To address this issue, we propose a novel, fully automatic, way to generate realistic occlusion pairs by exploiting the recent achievements in object segmentation in He et al. (2017). This results are high-fidelity occlusion pairs, where the background of the original image is preserved and the generated occlusion is more realistic. Additionally, we created a massive CG graphics generated dataset[1], in which we artificially created a large collection of occluded pedestrians. Additionally, we recovered from the game engine their attributes by manually annotating just the models. To our knowledge, this is the first CG dataset for the purpose of de-occluding people having a set of annotated person attributes (e.g. sex, hair color, dress style, etc.).

To summarize, our contributions are threefold:

- We propose a novel generative adversarial network that is able to solve occlusions in pedestrian images by hallucinating the missing parts while keeping both the appearance and the background coherent;
- We devise a new way for synthetically generating occlusion pairs that result in more realistic images when compared to other methods previously employed, also by creating a huge CG dataset;
- We propose a method for conditioning the occluded body part restoration on pedestrian attributes and consequently improving the generation process.

We show by experiments that the design of a conditional GAN that is aware of the attributes can acceptable hallucinate pedestrian and experimentally demonstrate that this information is helpful in guiding the generation process. Results are interesting in terms of very high accuracy, outperforming other previous methods. We believe that our method could be useful in many computer vision systems, from surveillance, automotive to human behavior understanding tasks.

[1] Leveraging on the highly photo-realistic graphics of GTAV video-game.

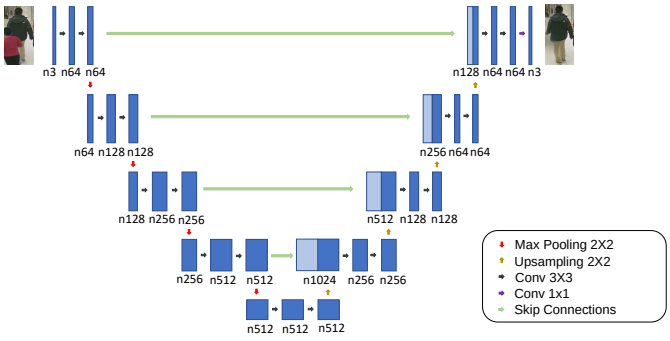

Fig. 2. Architecture of our generator network with corresponding number of feature maps (n), we always use 1 as stride size.

2. Related Works

Generative image modeling with deep learning techniques has received lots of attention in recent years. Works on this field can be split into two categories. The first line of works follows the unsupervised setup. Here, the variational autoencoders (VAE) proposed by Rezende et al. (2014) and Kingma and Welling (2013) are the first popular methods which apply a re-parameterization trick to maximize the lower bound of the data likelihood. The most popular methods are indeed generative adversarial networks (GAN) of Goodfellow et al. (2014) and Radford et al. (2015), which simultaneously learn a generator network to generate image samples and a discriminator network to discriminate generated samples from real ones. GANs are capable of generating sharp images by exploiting the adversarial loss instead of more canonical losses such L1 or L2.

The second group of works produce images conditioned on either categories, attributes, labels, images or texts. Yan et al. (2016) proposed a Conditional Variational Autoencoder (CVAE) to achieve an image generation conditioned on attributes. On the other hand, Mirza and Osindero (2014) proposed conditional GANs (CGAN) where both the generator and the discriminator are conditioned on extra information to perform category specific image generation. Lassner et al. (2017) generated people in clothing, by conditioning on the fine-grained body part segments. Reed et al. (2016a) proposed a novel deep architecture and GAN formulation to effectively translating visual concepts from characters to pixels, by adding textual information to both generator and discriminator. They also further investigated the use of additional location, keypoints or segmentation information to generate images in Reed et al. (2016c) and Reed et al. (2016b). With only these visual hint as condition and in contrast to our explicit condition on the occluded image, the control exerted over the image generation procedure is still abstract. Many works perform a conditioning over image generation not only on labels or texts, but also on images. Zhao et al. (2017) generated multi-view cloth images from only a single view input by proposing a new image generation model that combines the strengths of the variational inference and the GAN framework. Chen and Grauman (2014) tackled the unseen view inference by casting the problem in terms of tensor completion, and adapt a factorization approach to accommodate single-view images. Isola et al. (2017) provides a general purpose architecture that is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. Yang et al. (2015), Huang et al. (2017), Yim et al. (2015), Ghodrati et al. (2015) addressed the task of face image generation conditioned on a reference image and a specific face viewpoint. Finnaly Yang et al. (2017); Yeh et al. (2017); Pathak et al. (2016); Wang et al. (2018a) tackled the task of image inpainting where large missing regions have to be filled based on the available visual data. Our work can be seen as a particular case of inpainting, where the portion of the image to inpaint is not known a priori.

3. Method

The goal of our work is to reconstruct occluded body part of pedestrians in different surveillance scenarios. Given an image of an occluded pedestrian as the network input, we aim at removing the obstructions and replacing them with body parts that could likely belong to the occluded person. Note that, differently from the task of inpainting, we don't want to guide the network with the information about what portion of the image we want to remove and complete. For this purpose, we want to learn an image transformation between pairs of occluded images I_{occ} and not occluded images I_{GT}. To achieve this, we train a generator network G as a feed-forward CNN G_{θ_g} with parameters θ_g. For N training pairs images (I_{occ}, I_{GT}) we solve:

$$\hat{\theta}_g = \arg\min_{\theta_g} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{total}\left(G_{\theta_g}\left(I^n_{occ}\right), I^n_{GT}\right) \quad (1)$$

Here $\hat{\theta}_g$ is obtained by minimizing the loss function $\mathcal{L}_{total}$ described in the next subsection. Differently from Goodfellow et al. (2014); Radford et al. (2015), our generator network takes an image as input and instead of a random noise vector, as we want to be deterministic in terms of generated image (e.g. the same person should be always de-occluded in the same manner). As a result, our generator network learns a mapping from observed images I_{occ} to output image I_{gen}. This differs also from Isola et al. (2017); Mirza and Osindero (2014) which use random noise alongside with the input image.

Following Goodfellow et al. (2014), we further define the discriminator network D_{θ_d} with parameters θ_d, that we train alongside G_{θ_g} with the aim of solving the adversarial min-max problem:

$$\min_G \max_D \mathbb{E}_{I_{GT} \sim p_{data}(I_{GT})}[\log D\left(I_{GT}\right)] + \mathbb{E}_{I_{occ} \sim p_{gen}(I_{occ})}[\log 1 - D\left(G\left(I_{occ}\right)\right)] \quad (2)$$

where $D(I_{GT})$ is the probability of I_{GT} being a "real" image while $1 - D(G(I_{occ}))$ is the probability of $G(I_{occ})$ being a "fake" image. The min-max formulation force the generator network G to fool the discriminator network D, which is adversarially trained to distinguish between generated "fake" images and "real" ones. Thanks to this approach, we obtain a generator network G capable of learning solutions that are similar to not occluded images thus indistinguishable by the discriminator network D. Note also that, differently from Isola et al. (2017), we

Table 1. Classification performances of our ResNet-101 on RAP dataset

Method	mA	Accuracy	Precision	Recall	F1
ACN Sudowe et al. (2015)	69.66	62.61	**80.12**	72.26	75.98
DeepMAR Li et al. (2015)	73.79	62.02	74.92	76.21	75.56
DeepMAR* Li et al. (2016)	74.44	63.67	76.53	77.47	77.00
HP-Net Liu et al. (2017)	76.12	65.39	77.33	78.79	78.05
ACN-Res50 Fabbri et al. (2017)	**79.73**	64.13	76.96	78.72	77.83
Ours	78,46	**65,81**	77.81	**79.13**	**78.46**

Table 2. Detailed comparison between various pedestrian attribute datasets

Dataset	# Scenes	# Samples	# Attributes	Min. Resolution	Max. Resolution
PETA Deng et al. (2014)	-	19,000	61(+4)	17×39	169×365
RAP Li et al. (2016)	26	41,585	69(+3)	36×92	344×554
PA-100K Liu et al. (2017)	598	100,000	26	50×100	758×454
AiC	512	125,000	24	36×87	533×1080

do not concatenate input images I_{occ} to the "fake" images I_{gen} or to the "real" images I_{GT} as discriminator input.

Generator Network. Our generator structure differs from those presented in Radford et al. (2015) and Fabbri et al. (2017): following Ronneberger et al. (2015) and Isola et al. (2017) we propose the "U-Net" like architecture depicted in Fig. 2. In particular, the structure of our network slightly differs from the one described in Ronneberger et al. (2015) and Isola et al. (2017). The network is composed by 4 down-sampling blocks and a specular number of up-sampling components. Each down-sampling block consists of 2 convolution layers with a 3×3 kernel. Every convolutional layer is followed by a batch normalization and a leaky ReLU activation. Finally, each block has a max-pooling layer with stride of 2. The up-sampling part has a very similar but overturned structure, where each block is composed by an up-sampling layer of stride 2. After that, each block is equipped with 2 convolution layers with a 3×3 kernel. The last block has an additional 1×1 kernel convolutional layer which is employed to reach the desired number of channels: 3 RGB channels in our case. A *tanh* has been used as final activation. We additionally inserted skip connections between mirrored layers, in the down-sampling and up-sampling streams, in order to shuttle low-level information between input and output directly across the network. Eventually, padding is added to avoid cropping the feature maps coming from the skip connections and concatenate them directly to the up-sampling blocks outputs. Roughly speaking, our task can be seen as a particular case of image-to-image translation, where a mapping is performed between the input image and the output image. Additionally, for the specific problem we are considering, input and output share the same underlying structure despite differing in superficial appearance. Therefore, a rough alignment is present between the two images. In fact, all the non-occluded parts that are visible in the input images must be transferred to the output with no alterations. The structure of the U-Net lends itself optimally to our task, and the skip connections are fundamental for the conservation of the non-occluded image content. In this way, useful low-level information is not lost during the encoding passage: by leveraging this kind of information, we are able to maintain the appearance of visible parts in the image.

Discriminator Network. The discriminator, instead, aims to determine if an image is true or if it has been generated. In particular, the structure is similar to the one in Radford et al. (2015), composed by 4 convolutional layers with kernel size 5×5. The resulting features are followed by one sigmoid activation function in order to obtain a probability for the classification problem. We use batch normalization before every Leaky ReLU activation, except for the first layer.

3.1. Loss Function

The definition of the loss function $\mathcal{L}_{total}$ is crucial for the effectiveness of our generator network. We propose the following loss formulation, composed by a weighted combination of three components:

$$\mathcal{L}_{total} = \overbrace{\underbrace{\mathcal{L}_{adv}}_{\text{adver. loss}} + \underbrace{\lambda_1 \cdot \mathcal{L}_{vgg}}_{\text{cont. loss}} + \underbrace{\lambda_2 \cdot \mathcal{L}_{atr}}_{\text{attr. loss}}}^{\text{total loss}} \tag{3}$$

The first term of Eq. (3) is the adversarial loss $\mathcal{L}_{adv}$. This component encourages the generator network G to generate images belonging to the not occluded domain of pedestrians by fooling the discriminator network D:

$$\mathcal{L}_{adv} = \mathcal{L}_{bce}\left(D\left(G\left(I_{occ}\right)\right), 1\right) \tag{4}$$

where $D(G(I_{occ}))$ is the probability that $G(I_{occ})$ is classified as "real" by the discriminator network. As a possible drawback, the images produced by the generator network G are forced to be realistic thanks to the discriminator network D, but they can be unrelated with the original input. For instance, the output could be a plausible image of a pedestrian displaying a very different aspect with respect to the input image. Thus, is essential to mix the adversarial loss $\mathcal{L}_{adv}$ with an additional loss, such as

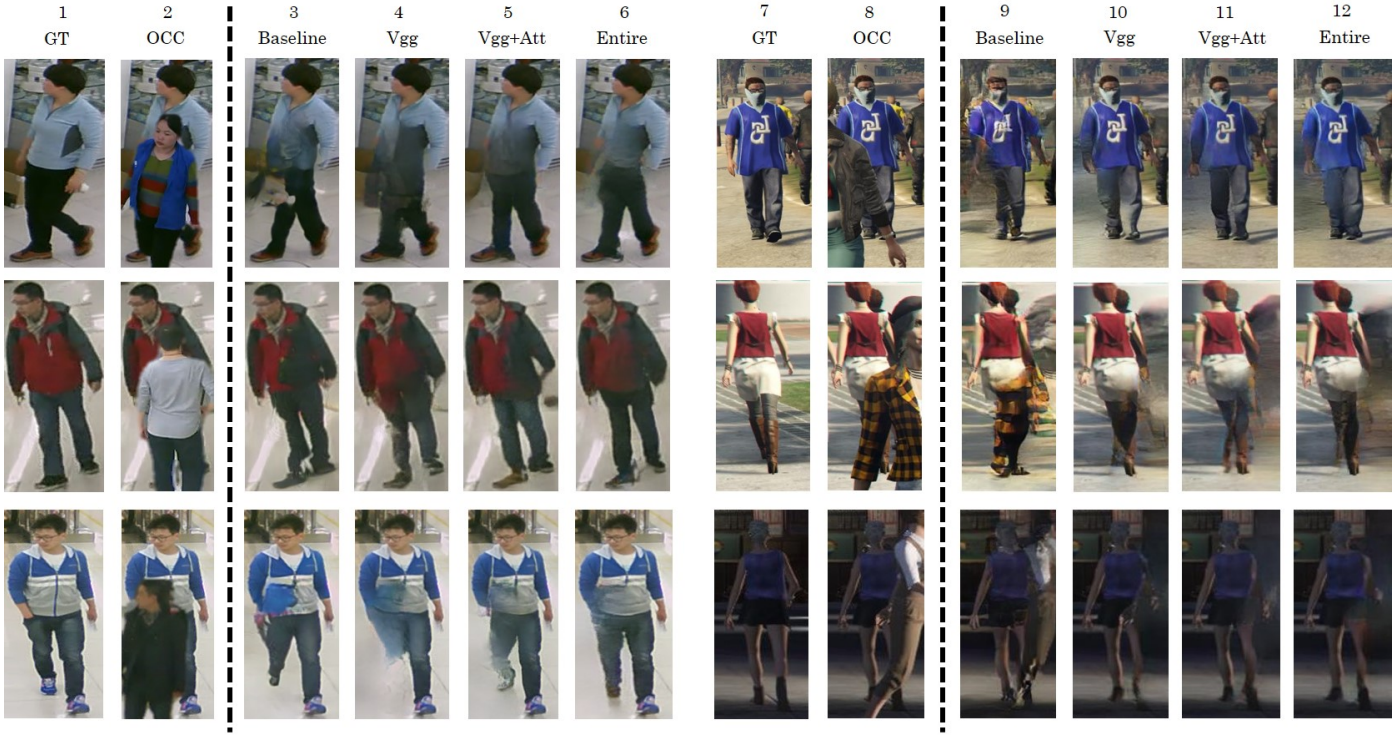

Fig. 3. Qualitative results based on the ablation study on RAP dataset (leftmost) and AiC dataset (rightmost). GT columns indicate ground truth images while in the OCC columns are presented the input occluded images. Columns 3 and 9 are the outputs of our baseline, where adversarial loss and MSE are used. Columns 4 and 10 represents results of the VGG loss. On 5 and 11 we have results of experiments using all the 3 losses combined: adversarial loss, VGG loss and attribute loss. Finally, columns 6 and 12 show results where attributes are injected as input to the network.

L1 or L2, that evaluate the per-pixel distance between the generated and the ground truth image. Usually, training a network using such losses leads to reasonable solutions. However, the outputs appear blurred with lack of high frequency details.

A possible solution for generating sharper results is to adopt a different content loss, like the perceptual loss introduced by Johnson et al. (2016) and used also in Ledig et al. (2017) or deblurring problems as in Kupyn et al. (2017):

$$\mathcal{L}_{vgg(i,j)} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I_o)_{x,y} - \phi_{i,j}(I_{gen})_{x,y}\right)^2 \quad (5)$$

where $W_{i,j}$ and $H_{i,j}$ are the dimensions of the feature maps $\phi_{i,j}$ obtained by the j-th convolution after ReLU activation and before the i-th max-pooling layer within the VGG16 network, pre-trained on ImageNet in Deng et al. (2009) as done by Johnson et al. (2016).

The $\mathcal{L}_{vgg}$ that we employed in our work is based on the sum of different intermediate layers of VGG16:

$$\mathcal{L}_{vgg} = \sum_{i,j \in L} \mathcal{L}_{vgg(i,j)} \quad (6)$$

where $\mathcal{L}_{vgg(i,j)}$ is taken from eq. 5 and L is the set of used activations. Rather than encouraging the pixels of the output image I_{gen} to exactly match the pixels of the target image I_{GT}, we instead encourage them to have similar feature representations as computed by the VGG16 network. As demonstrated in Johnson et al. (2016) and Mahendran and Vedaldi (2015), minimizing the content loss for higher layers do not preserve color and textures. As we reconstruct from early layers, instead, images tend to be perceptually similar to the target image I_{GT} in terms of color and texture. For this reason, we adopted early layers for the content loss objective.

Since our main purpose is not limited to naively restore the occluded parts of pedestrians, but also to maintain and highlight their attributes, we introduced an additional loss component $\mathcal{L}_{atr}$ of Eq. (3). Like for the perceptual loss $\mathcal{L}_{vgg}$, we used a classification network as loss function. More precisely we adapted ResNet-101 by He et al. (2016), pre-trained on ImageNet, to the task of multi-attribute classification. More precisely, we replaced the last two layers (the average pooling and the last fully connected layer) in order to fit the desired input shapes. Differently from the VGG loss, with this attribute loss, we work on a higher level of abstraction, forcing the generator network to produce images that exhibit characteristics coherent with the attributes of the person. In this case, we didn't use the euclidean distance as loss, but a weighted binary cross entropy:

$$\mathcal{L}_{atr} = -\sum_{i=1}^{A} \exp(1 - r_i) \cdot \left(y_i \cdot \log\left(\psi_i(I_{gen})\right)\right) + \exp(r_i) \cdot (1 - y_i) \cdot \log\left(1 - \psi_i(I_{gen})\right). \quad (7)$$

Here, A is the number of attributes classified by the ResNet-101, r_i is the positive ratio of i-th attribute. ψ is the output of our attribute classification network and y_i is the i-th ground truth label.

Fig. 4. Examples from the AiC dataset exhibiting its variety in viewpoints, illuminations and scenarios.

3.2. Training Details

We trained our GAN with 320×128 resized input images while simultaneously providing the target image in order to compute the supervised loss. We adopted the standard approach in Goodfellow et al. (2014) to optimize the networks alternating gradient descent updates between the generator and the discriminator with $K = 1$. Data augmentation is performed by randomly flipping the images horizontally. We used mini-batch SGD applying the Adam solver with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and learning rate $2 \cdot 10^{-4}$. In our experiments we chose a λ_1 value of 10 and λ_2 value of 5 of Equation 3 and a batch size of 20. Each training is performed using a Titan Xp GPU.

4. Datasets

We evaluated our method on RAP dataset from Li et al. (2016), comparing state-of-the-art methods and performing the ablation study over each loss employed. In addition, we further propose a new large-scale computer-graphics dataset AiC for pedestrian attribute recognition in crowded scenes. Differently from existing publicly available datasets, AiC is mainly focused on occlusion events.

4.1. RAP Dataset

RAP from Li et al. (2016) is a very richly annotated dataset with 41,585 pedestrian samples, each of which is labeled with 72 attributes as well as viewpoints, occlusions and body parts information. In order to evaluate our method, we corrupted the dataset with occlusions. Differently from Fabbri et al. (2017), where obstructions are created by cutting parts of images according to regular geometric shapes, we have adopted a more sophisticated approach that has led us to more realistic results. By exploiting the state-of-the-art performances of Mask R-CNN He et al. (2017), pre-trained on the COCO Dataset in Lin et al. (2014), we produced segmentation masks for each person in the RAP dataset. The computed silhouettes were then used to crop people's shapes from the dataset. Those crops are then used to reproduce the occlusions by simply randomly overlapping them to each image sample of RAP dataset. In addition, to reduce the visual gap between the original image and the overlapped person, we performed a Gaussian blurring. However, this is not applied to the whole image but only to the area given by the difference between an expansion and an erosion of the segmentation mask of the overlapping image. The only constraint that we have introduced is that the overhead person must not occupy 1/7 of the top part of the starting image. Each image is computed as follows:

$$I_{occ} = I_{GT^1} \odot \neg\alpha\left(\beta\left(I_{GT^2}\right)\right) + \alpha\left(\beta\left(I_{GT^2}\right) \odot I_{GT^2}\right) \qquad (8)$$

where $\beta(I_{GT^2})$ is the binary mask generated using Mask R-CNN and morphology operations. α is a function used to translate the overlap section randomly over the destination image I_{GT^1}. The dataset is already organized in 5 random splits. Each of which contains 33,268 images for training and 8,317 for testing. Due to the unbalanced distribution of attributes in RAP we selected the 51 attributes that have the positive example ratio in the dataset higher than 0.01.

4.2. AiC Dataset

Most of the publicly available pedestrian attribute datasets like RAP in Li et al. (2016); PETA in Deng et al. (2014);

Table 3. Ablation study results on RAP dataset

Method	mean Accuracy	Accuracy	Precision	Recall	F1	SSIM	PSNR
Occlusion	65.74	51.05	68.71	64.35	66.46	0.7079	14.62
Baseline	70.75	56.54	70.57	71.79	71.17	0.7853	20.35
VGG loss	72.42	58.83	72.53	73.53	73.02	**0.8181**	**20.89**
VGG and attr. loss	72.22	59.59	73.47	73.76	73.61	0.8143	20.68
VGG and attr. loss (+ input attr.)	**81.16**	**74.76**	**84.23**	**85.63**	**84.92**	0.8151	20.72
GT data	78,46	65,81	77.81	79.13	78.46	-	-

Table 4. Ablation study results on Aic dataset

Method	mean Accuracy	Accuracy	Precision	Recall	F1	SSIM	PSNR
Occlusion	73.09	65.00	75.53	75.30	75.41	0.6125	19.54
Baseline	74.00	66.41	76.12	76.63	76.37	0.6028	21.08
VGG loss	79.04	71.98	80.10	81.21	80.65	0.7019	**23.01**
VGG and attr. loss	82.30	76.45	83.66	84.41	84.03	**0.7048**	22.90
VGG and attr. loss (+ input attr.)	**91.93**	**88.73**	**92.40**	93.36	**92.88**	0.7021	22.87
GT data	91.59	88.19	91.67	**93.50**	92.57	-	-

PA-100K in Liu et al. (2017) does not contemplate occlusion events. They only provide samples of full visible people, completely ignoring crowded situations of pedestrians occluding each other (which is indeed common in urban scenarios). To overcome this limitation, we propose the Attributes in Crowd dataset, a novel synthetic dataset for people attribute recognition in presence of strong occlusions. AiC features 125,000 samples (100,000 for training and 25,000 for testing), each of which is automatically labeled with information concerning sex, age etc. Each of the 24 attributes is present at least in a 15% of samples which highlight a good balance in terms of labels. The collected samples feature a vast number of different body poses, in several urban scenarios with varying illumination conditions and viewpoints. Skeleton joints are also available for each identity. Joints are additionally labeled with an occlusion flag which tells if the specific body part is directly visible from the camera point of view. Moreover, each image sample has his vanilla version where each obstacle is removed from the image. Thus, for each occluded pedestrian, we know exactly how it really is behind the occlusion (this is obviously not obtainable in real environments). Fig. 4 exhibits some examples of the dataset. To foster future research on this topic, the dataset will be publicly released upon publication. AiC was created by exploiting the highly photo-realistic video game *Grand Theft Auto V* developed by *Rockstar North*.

5. Experimental Results

In this section we provide details about the metrics adopted, followed by a detailed ablation study that presents qualitative and quantitative results for three different combinations of losses (that we added to the adversarial loss): MSE loss, VGG loss and a combination of VGG loss and attribute loss. We also investigate how the information about the attributes of a person can enhance the quality of the produced images. Finally, we compare our method with the most related works of Isola et al. (2017) and Fabbri et al. (2017).

5.1. Evaluation Metrics

Evaluating the quality of synthesized images is an open and challenging problem as stated in Salimans et al. (2016). Traditional metrics such as per-pixel MSE do not estimate joint statistics of the result, and therefore do not extrapolate the full structure of the result. In order to more holistically evaluate the visual quality of our results, we employed two tactics. Firstly, we compared the performance of the proposed model through metrics directly calculated over the reconstructed images. Specifically, we adopted the structural similarity SSIM and the peak signal-to-noise ratio PSNR. Secondly, we measured the capability of the proposed network of being able to preserve original attributes, like gender, hairstyle or wearing jacket, by exploiting the ResNet-101 of He et al. (2016) network trained on the task of multi-attribute classification. Thus, following Li et al. (2016), Fabbri et al. (2017) and Liu et al. (2017), we provide five evaluation metrics for the attribute classification task, namely mean Accuracy, Accuracy, Precision, Recall and F1.

ResNet-101 Classification Network. We trained the network with 320×128 resized images with Adam as optimizer and learning rate set to $2 \cdot 10^{-4}$. In Table 1 a comparison on the classification task with other state-of-the-art networks on RAP dataset is presented. The same network is trained independently using RAP and AiC, in order to provide reliable metrics for each dataset.

5.2. Ablation Study

As previously stated, we investigated three loss combinations in order to clarify and highlight the solutions adopted in our work:

Fig. 5. Qualitative comparison with state-of-the-art approaches: results are presented for both RAP (leftmost) and AiC (rightmost). GT columns indicate ground truth images while in the OCC columns are presented the input occluded images. Columns 1 and 4 are the images recovered by Pix2Pix in Isola et al. (2017). On 2 and 5 are presented results obtained from the method used in Fabbri et al. (2017). The last two columns, 3 and 5, show our best approach output.

- *Baseline*: the Baseline architecture uses, in conjunction with the adversarial loss, the MSE loss as content loss;
- *VGG loss*: differently from the Baseline, we replaced the MSE loss with the VGG loss. The layers (1,2), (2,2), (3,3) and (4,3) are chosen as the set L of activations on Eq. 6. In Eq. 3 we set λ_1 to 10 and λ_2 to 0;
- *VGG loss + Attr. loss*: in this case all the three losses are employed. The VGG loss always refers to the same four activation layers. The Attribute loss is computed between the output of the ResNet-101 classification network computed on the generated images and the ground truth labels provided by the datasets. In Eq. 3 we set λ_1 and λ_2 to 10 and 5 respectively. Note that we did not use all the available attributes of RAP dataset, but only the first 51 for the reason explained in section 5.1. For AiC dataset, instead, we used all the available attributes.

In order to further investigate how some additional information about the attributes can improve the restoration process, we performed a further experiment where attributes are fed as input to the network, alongside with the occluded image:

- *Entire*: in this setup we adopted both the VGG loss and the Attribute loss, alongside with the adversarial loss. Differently from our main method, attributes are injected directly to the main flow of the generator network. Specifically, the attribute vector of the occluded pedestrian is fed to a fully connected layer in order to produce a feature vector that is reshaped to match the bottleneck dimension of our generator network.

Fig. 3 shows some qualitative results. The baseline performs considerably worse than the other setups, not being able to completely remove the occlusions on AiC (column 9 of Fig. 3). This is probably due to the fact that AiC is a more challenging dataset compared to our corrupted version of RAP. For the same reason, RAP results are overall more appealing than the ones of AiC. Moreover, no substantial difference appears between the other setups, highlighting the fact that the VGG loss is the main component that guides the network to produce high-quality results.

Table 3 and Table 4 provide quantitative results for RAP and AiC respectively. From the tables it emerges that, despite being visually indistinguishable, the images obtained from the three setups (VGG loss, VGG and Attribute loss and Entire) produce very different results in terms of attribute metrics. In particular, the difference between the VGG loss and the VGG loss with Attribute loss on AiC dataset differs by about 3 points. Moreover, by injecting the generator network with information concerning attributes, we obtain attribute metrics remarkably higher compared to the upper bound of the ground truth images. The generator network, by restoring the occluded images, is able to produce an output that has enhanced attribute characteristics.

Table 5. Comparison with the state-of-the-art method on RAP dataset

Method	mA	Accuracy	Precision	Recall	F1	SSIM	PSNR
Occlusion	65.74	51.05	68.71	64.35	66.46	0.7079	14.62
Pix2Pix Isola et al. (2017)	69.53	52.00	64.95	70.05	67.40	0.7172	17.94
Fabbri et al. (2017)	66.00	51.40	65.59	67.97	66.76	0.6758	18.47
Ours	**72.22**	**59.59**	**73.47**	**73.76**	**73.61**	**0.8143**	**20.68**

Table 6. Comparison with the state-of-the-art method on AiC dataset

Method	mA	Accuracy	Precision	Recall	F1	SSIM	PSNR
Occlusion	73.09	65.00	75.53	75.30	75.41	0.6125	19.54
Pix2Pix Isola et al. (2017)	69.52	60.69	72.12	71.86	72.00	0.6318	20.91
Fabbri et al. (2017)	73.10	64.79	75.48	75.01	75.24	0.5949	21.23
Ours	**82.30**	**76.45**	**83.66**	**84.41**	**84.03**	**0.7048**	**22.90**

5.3. Comparison With Previous Works

Since our task of de-occlusion is novel, there is no direct comparison work. So, to compare the results of our network, in addition to Fabbri et al. (2017), we applied the pix2pix architecture in Isola et al. (2017) to our task. In Table 5 and Table 6 can be shown that our network perform favourably for each metric, both for RAP and AiC datasets.

From Fig. 5 it emerges that our method, despite not using attention mechanisms, is able to detect and to remove the occlusion, with no external additional information. Furthermore, differently from Fabbri et al. (2017) and Isola et al. (2017), our method learns to transfer with no alterations the portion of images that are not occluded. Finally, Fig. 6 depicts some failure cases of our method that display the challenge of strong occlusions.

Fig. 6. Some failure cases on RAP (leftmost) and AiC (rightmost). From this images, it is possible to see that, in cases of strong occlusions, complete restoration remains a difficult task. Moreover, also with particular background conditions, as we can see in the first image triplet, the network is not able to perform a reliable reconstruction.

6. Conclusions

In this work we presented the use of GANs for image enhancing in people attributes classification. Our generator network have been designed to overcome a common problem in surveillance scenarios, namely people occlusion. Experiments have shown that jointly enhancing images before feeding them to an attribute classification network can improve the results even when input images is affected by this issue. We find this line of work can foster research about the problem of attribute classification in surveillance contexts where camera resolution and positioning cannot be neglected.

References

Chen, C.Y., Grauman, K., 2014. Inferring unseen views of people, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2003–2010.

Coppi, D., Calderara, S., Cucchiara, R., 2016. Transductive people tracking in unconstrained surveillance. IEEE Transactions on Circuits and Systems for Video Technology 26, 762–775. doi:10.1109/TCSVT.2015.2416555.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09.

Deng, Y., Luo, P., Loy, C.C., Tang, X., 2014. Pedestrian attribute recognition at far distance, in: Proceedings of the 22Nd ACM International Conference on Multimedia.

Fabbri, M., Calderara, S., Cucchiara, R., 2017. Generative adversarial models for people attribute recognition in surveillance, in: Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on, IEEE. pp. 1–6.

Ghodrati, A., Jia, X., Pedersoli, M., Tuytelaars, T., 2015. Towards automatic image editing: Learning to see another you. arXiv preprint arXiv:1511.08446 .

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR, IEEE Computer Society. pp. 770–778.

Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: A survey. Image and Vision Computing 60, 4 – 21. Regularization Techniques for High-Dimensional Data Analysis.

Huang, R., Zhang, S., Li, T., He, R., et al., 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. arXiv preprint arXiv:1704.04086 .

Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 5967–5976. doi:10.1109/CVPR.2017.632.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II, pp. 694–711.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 .

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J., 2017. Deblurgan: Blind motion deblurring using conditional adversarial networks. CoRR abs/1711.07064.

Lassner, C., Pons-Moll, G., Gehler, P.V., 2017. A generative model of people in clothing, in: Proceedings of the IEEE International Conference on Computer Vision.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 105–114. doi:10.1109/CVPR.2017.19.

Li, D., Chen, X., Huang, K., 2015. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) , 111–115.

Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K., 2016. A richly annotated dataset for pedestrian attribute recognition. preprint arXiv:1603.07054 .

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yan, J., Wang, X., 2017. Hydraplus-net: Attentive deep features for pedestrian analysis, in: Proceedings of the IEEE international conference on computer vision, pp. 1–9.

Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K.M., Zhong, Y., 2017. Person re-identification by unsupervised video matching. Pattern Recognition 65, 197 – 210.

Mahendran, A., Vedaldi, A., 2015. Understanding deep image representations by inverting them, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5188–5196.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 .

Ouyang, W., Zeng, X., Wang, X., 2016. Partial occlusion handling in pedestrian detection with a deep model. IEEE Transactions on Circuits and Systems for Video Technology 26, 2123–2137.

Pan, J., Hu, B., 2007. Robust occlusion handling in object tracking, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. doi:10.1109/CVPR.2007.383453.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 .

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016a. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 .

Reed, S., van den Oord, A., Kalchbrenner, N., Bapst, V., Botvinick, M., de Freitas, N., 2016b. Generating interpretable images with controllable structure .

Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H., 2016c. Learning what and where to draw, in: Advances in Neural Information Processing Systems, pp. 217–225.

Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082 .

Riza Alp Guler, Natalia Neverova, I.K., 2018. Densepose: Dense human pose estimation in the wild.

Ronneberger, O., P.Fischer, Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 234–241.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans, in: Advances in Neural Information Processing Systems, pp. 2234–2242.

Subramaniam, A., Chatterjee, M., Mittal, A., 2016. Deep neural networks with inexact matching for person re-identification, in: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 2667–2675.

Sudowe, P., Spitzer, H., Leibe, B., 2015. Person attribute recognition with a jointly-trained holistic cnn model., in: ICCV Workshops, IEEE Computer Society. pp. 329–337.

op het Veld, R.M., Wijnhoven, R., Bondarev, Y., et al., 2015. Detection and handling of occlusion in an object detection system, in: Video Surveillance and Transportation Imaging Applications 2015, International Society for Optics and Photonics. p. 94070N.

Wang, C., Xu, C., Wang, C., Tao, D., 2018a. Perceptual adversarial networks for image-to-image transformation. IEEE Transactions on Image Processing 27, 4066–4079.

Wang, X., Hou, Z., Yu, W., Pu, L., Jin, Z., Qin, X., 2018b. Robust occlusion-aware part-based visual tracking with object scale adaptation. Pattern Recognition 81, 456 – 470.

Yan, X., Yang, J., Sohn, K., Lee, H., 2016. Attribute2image: Conditional image generation from visual attributes, in: European Conference on Computer Vision, Springer. pp. 776–791.

Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H., 2017. High-resolution image inpainting using multi-scale neural patch synthesis, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 3.

Yang, J., Reed, S.E., Yang, M.H., Lee, H., 2015. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis, in: Advances in Neural Information Processing Systems, pp. 1099–1107.

Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N., 2017. Semantic image inpainting with deep generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5485–5493.

Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J., 2015. Rotating your face using multi-task deep neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 676–684.

Zhao, B., Wu, X., Cheng, Z.Q., Liu, H., Jie, Z., Feng, J., 2017. Multi-view image generation from a single-view. arXiv preprint arXiv:1704.04886 .

Zhuo, J., Chen, Z., Lai, J., Wang, G., 2018. Occluded person re-identification. CoRR abs/1804.02792.